---
title: "Deployment Task"
author: "Jacob López Minguela, Antonio Martí Mora"
date: "2023-05-04"
output: html_document
---

1. One dimensional Partial Dependence Plot.

```{r, warning=FALSE}
library(readr)
library(pROC)
library(dplyr)
library(ggplot2)
library(randomForest)
library(plotly)
library(grid)
library(gridExtra)
library(fairness)
library(caret)
library(mlbench)
library(partykit)
library(pre)
library(pdp)
```

Apply PDP to the regression example of predicting bike rentals. Fit a random forest approximation for the prediction of bike rentals (cnt). Use the partial dependence plot to visualize the relationships the model learned. Use the slides shown in class as model.

Analyse the influence of days since 2011, temperature, humidity and wind speed on the predicted bike counts.

```{r}
bike <- read.csv("day.csv")

# One-hot encoding for season
bike <- bike %>%
  mutate(SPRING = ifelse(season == 1, 1, 0),
         SUMMER = ifelse(season == 2, 1, 0),
         FALL = ifelse(season == 3, 1, 0),
         WINTER = ifelse(season == 4,1,0))

# Create the MISTY and RAIN features
bike$MISTY <- ifelse(bike$weathersit == 2, 1, 0)
bike$RAIN <- ifelse(bike$weathersit %in% c(3, 4), 1, 0)

# Denormalization
bike$temp <- bike$temp * 41
bike$hum <- bike$hum * 100
bike$windspeed <- bike$windspeed * 67

# Create days_since_2011 feature
bike$days_since_2011 <- as.integer(difftime(as.Date(bike$dteday), as.Date("2011-01-01"),
units = "days"))


# Random Forest
random_forest <- randomForest(cnt ~ workingday + holiday + SUMMER + FALL + WINTER +
MISTY + RAIN + temp + hum + windspeed + days_since_2011, data = bike, ntree = 400, mtry
= 3)

random_forest
summary(random_forest)
```

```
# PDP
days_2011 <- partial(random_forest, pred.var = "days_since_2011", plot = TRUE,
plot.engine = "ggplot2")

temperature <- partial(random_forest, pred.var = "temp", plot = TRUE, plot.engine =
"ggplot2")

humidity <- partial(random_forest, pred.var = "hum", plot = TRUE, plot.engine =
"ggplot2")

windspeed <- partial(random_forest, pred.var = "windspeed", plot = TRUE, plot.engine =
"ggplot2")

days_2011
temperature
humidity
windspeed
```
Based on the partial dependence plots we have created following up the slides shown in class, we can analyse the influence of days since 2011, temperature, humidity and wind speed on the predicted bike counts:

• days_since_2011: The rental bike count has been on the rise since 2011, and we expect this due to the company's ability to grow up, resulting in more bikes being available for rental. The more bikes for rental, the more clients.

• temperature: When the weather condition strings along with the day, the more people want to rent a bike, as we can see on the influence on temperature on the predicted bike counts. When temperatures are warm, between 20 and 25 degrees, people tend to rent bikes. Nevertheless, when the weather is cold, or even too hot, it's quite more difficult to find clients in order to rent a bike.

• humidity: The worse weather condition, the less expected rents. High values of humidity are related to bad weather conditions, including rain or even fog, where taking a bike is not the optimal plan to do. It is also dangerous due to the fact that the floor can skid. Therefore, we can observe the decrease in the predicted number of bike rentals when humidity is up to 65 %.

• windspeed: Higher wind speeds also lead to less bike rentals. As soon as the wind is significant and exceeds 10km/h, we can observe a progressive decrease in the predicted number of bike rentals. Wind is not optimal for a bike riding.

2. Bidimensional Partial Dependency Plot.

Generate a 2D Partial Dependency Plot with humidity and temperature to predict the number of bikes rented depending on those parameters.
Show the density distribution of both input features with the 2D plot as shown in the class slides.

```{r}

set.seed(150)
bike_sample <- bike %>% sample_n(200)

random_forest_bike <- randomForest(cnt ~workingday + holiday + SUMMER + FALL + WINTER +
MISTY + RAIN + temp + hum + windspeed + days_since_2011, data = bike_sample, ntree =
400, mtry = 3)

pdp_tem_hum <- partial(random_forest_bike, pred.var = c("temp", "hum"), plot = FALSE)

ggplot(pdp_tem_hum, aes(x = temp, y = hum, fill = yhat)) +
  geom_tile(aes(width = 3, height = 3)) +
  scale_fill_gradient(low = "white", high = "red") +
  geom_density2d() +
  theme_bw() +
  xlab("Temperature") +
  ylab("Humidity") +
  ggtitle("Bidimensional Partial Dependency Plot")
```

```

With this 2D Partial Dependency Plot, made by following up the slides shown in class, we can observe how the number of bikes rented depends on both temperature and humidity. With temperatures below 20 degrees and a humidity more than 70%, people refuse to rent bikes due to the inappropiate weather conditions of cold temperatures or extremely heat. When temperatures are warm, between 20 and 25 degrees, and humidity varies between 40% and 60%, it is the perfect climate situation for people to rent a bike (as we can see on the plot with the most colorful red).

## 3. PDP to explain the price of a house.

Apply the previous concepts to predict the price of a house from the database kc_house_data.csv. In this case, use again a random forest approximation for the prediction based on the features bedrooms, bathrooms, sqft_living, sqft_lot, floors and yr_built.

```{r}
house <- read.csv("kc_house_data.csv")

# Select random samples
set.seed(150)
house_sample <- house %>% sample_n(900)

# Fit random forest model
random_forest_house <- randomForest(price ~ bedrooms + bathrooms + sqft_living +
sqft_lot + floors + yr_built, data = house_sample, ntree = 400, mtry = 3)

# Generate partial dependence plot for selected features
bedrooms <- partial(random_forest_house, pred.var = c("bedrooms"), plot = TRUE,
plot.engine = "ggplot2")

bathrooms <- partial(random_forest_house, pred.var = c("bathrooms"), plot = TRUE,
plot.engine = "ggplot2")

sqft <- partial(random_forest_house, pred.var = c("sqft_living"), plot = TRUE,
plot.engine = "ggplot2")

floors <- partial(random_forest_house, pred.var = c("floors"), plot = TRUE, plot.engine
= "ggplot2")

bedrooms
bathrooms
sqft
floors

```

· bedrooms: The number of bedrooms does not generate the biggest impact on the price of a house.As we can see on the plots, there are no significant differences on house pricing between 2 and 4 bedrooms per house. Furthermore, houses with five bedrooms tend to be the cheapest. However, when we are talking about houses with 6 or more bedrooms, referred to kind of mansions, palaces or large villas, the price increases substantially.

· bathrooms: Same for bathrooms, the more bathrooms a house hold, the more the price increases. Number of bathrooms are also proportional to the size of the house.

· sqft_living: Looking at the graph, we can observe a correlation between the square footage of a house (sqft_living) and its price, which appears to follow a linear trend. As the size of the house in square feet increases, the estimated price of the house also increases in a linear fashion. However, once the square footage exceeds 7000, the prices of houses become extremely high, exceeding 2000000.
```

• floors: The number of floors in a house has a substantial impact on its price. It is evident that houses with three or more floors command a considerably higher price compared to those with fewer floors, with the price increasing exponentially with each additional floor. Conversely, houses with around two floors have a comparatively lower price, and the price tends to decrease as the number of floors decreases. Hence, houses with one or two floors are more economical in terms of price.